4-2004

Combining nearest neighbor classifiers versus cross-validation selection

Minhui Paik
Iowa State University, 100min@gmail.com

Yuhong Yang
Iowa State University

Combining Nearest Neighbor Classifiers Versus Cross-Validation Selection

Minhui Paik and Yuhong Yang
Department of Statistics
Iowa State University
Ames, IA, 50011

April 8, 2004

Abstract

Various discriminant methods have been applied for classification of tumors based on gene expression profiles, among which the nearest neighbor (NN) method was reported to perform relatively well. Usually cross-validation (CV) is used to select the neighbor size as well as the number of genes for the NN method. However, CV can perform poorly when there is considerable uncertainty in choosing the best candidate classifier. As an alternative to selecting a single "winner", in this work, we propose a weighting method to combine the multiple NN rules. Three gene expression data sets are used to compare its performance with CV methods. The results show that when the CV selection is unstable, the combined classifier performs much better.

1 Introduction

The availability of rich gene expression data opens new channels for obtaining valuable information regarding certain biological and medical questions. From the statistical point of view, the nature of such data presents numerous statistical questions on how to accurately extract information from the microarray experiments. In the specific context of tumor classification with gene expression data, Dudoit, Fridlyand and Speed (2002) listed three main statistical issues, (i) the identification of new/unknown tumor classes using gene expression profiles, (ii) the classification of malignancies into known classes, and (iii) the identification of "marker" genes that distinguish among the different tumor classes. Our concern in this work is on the classification accuracy of certain classification rules (i.e., the second issue above), focusing on tumor classification with gene expression data sets in our empirical investigation.

A number of classification methods have been used for gene expression data. For example, Golub *et al.* (1999) used a modified linear discriminant analysis to classify leukemia cancers. Other methods include nearest neighbor (Fix and Hodges (1951)), flexible discriminant analysis (Hastie *et al.* (2001)), shrunken centroid classifier (Tibshirani (2002)), CART (Breiman *et al.* (1984)), support vector machine (SVM) (Vapnik (1982, 1998)), bagging (Breiman (1996)), and boosting (Freund and Shapire (1997)). To compare the performance of the different approaches on gene expression data, Dudoit *et al.* (2002) used three data sets to empirically assess the merits of the competing methods in terms of classification error rate. Their results suggest that simple methods tend to do well or better than the more complicated alternatives. In particular, they found that the nearest neighbor (NN) method performed very well.

For a successful application of the NN method, one needs to choose appropriately the number of neighbors and also the set of feature variables. Model selection methods such as cross validation (CV) are frequently applied for that purpose. However, in recent years, it has been observed that when model selection uncertainty is high, combining or mixing the models/procedures instead can substantially improve prediction/estimation accuracy.

In this paper, we propose a combining method to mix the NN classifiers with different choices of the neighbor size and the feature variables. As will be seen, the empirical comparisons with CV selection based on the three gene expression data sets in Dudoit *et al.* (2002) clearly demonstrate the potential advantage of the combining approach. The results are consistent with earlier ones that compared model selection with model combining in other contexts (e.g., Yang (2003)).

Even though there have been a number of empirical studies on combining classifiers (including bagging and boosting), there is little investigation on when combining is better. This is a very important issue because it is not true that combining is always better and in fact, combining can perform very poorly. To address this issue, using the gene expression data sets, we examine the relationship between CV selection instability and the performance of combining classifiers relative to CV selection. It is seen that when the CV selection is unstable, combining performs better; and when CV selection is very stable, combining does not lead to a better accuracy and and it can even hurt the performance.

The rest of the paper is organized as follows. Section 2 reviews and discusses the NN method, cross-validation and model selection uncertainty. The combining method of this paper is proposed in Section 3. Section 4 briefly describes the data sets used in the empirical investigation. Section 5 explains the design of our empirical study. In Section 6, we report the results of the comparison of CV selections with our combining approach based on the three data sets and an effort is also made to understand when combining is better than selecting. Section 7 gives concluding remarks.

2 Nearest Neighbor classifiers and Cross-Validation

The nearest neighbor (NN) method is one of the simplest and well-known nonparametric classification methods. By the k-NN rule ($k \geq 1$), to classify a new case with the feature variable values known, one simply looks at the k nearest neighbors in the available data and the class label with the highest frequency wins. For defining neighbors, a distance or metric is usually taken. Not surprisingly, the performance of the NN method may depend heavily on the chosen distance and different distances work well on different data sets.

Given a set of feature variables, a key issue in the NN classification is the choice of the neighbor size k. Cover and Hart (1967) showed that even 1-nearest neighbor rule can do half as well as the Bayes rule (the optimal classifier) in terms of the classification error probability. Larger choices of k can often improve classification accuracy. Basically, a large choice of k reduces the variability of the classifier but at the same time increases the bias in the approximation of the conditional probability functions (i.e., the conditional probability of the response given the feature value). A good balance between the two competing directions is necessary for a high classification accuracy. It has been shown that with an appropriate choice of the neighbor size, NN classification is universally consistent (regardless of the underlying distribution of the response and the feature variables). It can also converge at the optimal rate for certain classes of conditional probability functions. Interested readers are referred to Devroye, Györfi and Lugosi (1996) for details and many related interesting results.

Feature selection is another important issue. In the context of gene expression data, a very large number of features (expression levels of all the studied genes, usually in thousands or tens of thousands) are available. Obviously, including many irrelevant genes or missing important ones can both substantially degrade the performance of the NN classifiers.

From the above discussion, for tumor classification with gene expression data, choosing the neighbor size and the genes (features) is a crucial aspect in the NN method.

Cross validation (CV) (e.g., Allen (1974), Stone (1974), Geisser (1975)) is a natural and commonly used approach to deal with model/procedure selection. Basically, a proportion of the data is used to assess the candidates (in our context, the candidates are nearest neighbor rules based on different choices of the numbers of neighbors and the genes) that are built on the rest of the data and the one with the best performance is taken. There are different versions of CV in terms of the implementation details.

Note that cross validation has been used for two very different purposes. One is to evaluate competing procedures by reserving part of the data as "future data" to assess the accuracy of the procedures (due to the fact that real "future" data are usually not available yet), and the other is to select a hyper-parameter (such as model size or smoothing parameters) within a classifier. See West *et al.* (2001), Ambroise and McLachlan (2002) and Speed (2003, Chapter 3) for more discussions on the use of CV for objective assessment of classifiers. Both usages of CV mentioned above are employed in this work.

Like other model selection methods, CV faces the problem of selection uncertainty or instability (e.g., Breiman (1996)). That is, when several models/procedures have CV values close to each other, the uncertainty in selection is substantial and usually a slight change of the data can cause a rather significant change of the classifier. This undesirable instability causes the classifier (or predictions/estimators in other contexts) to have a large variability and thus damage its classification accuracy.

Breiman (1996) proposed bootstrap aggregating, or bagging, to stabilize a classifier by averaging over a number of bootstrap samples. He reported significant improvements by bagging for unstable classifiers such as CART, but he regarded nearest neighbor rules stable. Empirical studies (Speed (2003, Chapter 3)) stated that bagging appeared to have little effect on k-NN classifiers.

Instead of going through model selection every time and then averaging, another approach to deal with model selection uncertainty is to weight the candidate models by appropriate sub-sampling and evaluation. Intuitively, when two classifiers are really close in terms of a selection criterion, appropriate weighting of them can be much better than an exaggerated 0 or 1 decision (i.e., selecting the "best" classifier). Yang (2000) proposed a weighting method to combine a list of candidate models/procedures for classification and derived its theoretical properties. In the context of NN classification, the same idea can be used to combine NN rules with different choices of the neighbor size and the number of feature variables. The objective of this paper is to provide a practically feasible weighting method for combining NN rules and compare its performance with selecting a single one based on cross-validations. As will be seen, when CV selection has a high uncertainty, combining the NN classifiers significantly improves the classification accuracy.

The combining method in this paper works for a general distance chosen for NN classification. For the empirical study, as in Dudoit *et al.* (2002), we standardize the feature variables to have sample mean 0 and variance 1 and consider the Euclidean distance between two mRNA samples, i.e., for $x, x^{'} \in R^p$,

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2 + \ldots + (x_p - x'_p)^2}.$$

Even though the choice of the distance for NN classification can play an important role, since our main interest in this paper is the comparison of selection versus combining, we will simply use the Euclidean distance throughout this work.

As mentioned earlier, for applying NN classification, we need to determine which feature variables to use. Since the number of genes is typically very large in micro-array data, considering all subsets is obviously computationally prohibitive. As in Dudoit *et al.* (2002), we consider only order selection, i.e., the feature variables are already ordered and then one only needs to select the number of variables (genes). In general, this can sometimes be a poor strategy because it is rarely the case that the importance of the feature variables can be pre-determined before an appropriate assessment of the candidate models. Note that, however, this is not much a concern for our comparison of CV selection and combining.

Now let us provide a little more detail on cross-validation for model/procedure selection. For K-fold cross-validation, one splits data into K roughly equal-sized parts; For the ith part, we fit each model using the other $K - 1$ parts of the data, and calculate the prediction error of the fitted model when predicting the ith part of data. Complete this for all i ($1 \leq i \leq K$) and find the average of prediction error. To reduce the effect of the order of the observations in data splitting, one may replicate this a number of times with a random permutation of the data. The model/procedure with the smallest average prediction error is selected. Two popular choices of K are 5 and 10 (4:1 scheme and 9:1 scheme). In addition, the case $K = n$ (n is the number of observations) is known as *leave-one-out* or *delete-one* cross-validation.

In this work, for selecting the neighbor size and the number of feature variables, 2-fold, 3-fold and *leave-one-out* cross-validations are used in the empirical study.

3 Combining nearest neighbor classifiers

Suppose that one observes $Z_i = (X_i, Y_i)$, $i = 1, ..., n$, independent copies of a random pair $Z = (X, Y)$ with class labels $Y \in \{1, ..., m\}$ (i.e., there are m classes) and the feature vector X (consisting of d feature variables) taking value in R^d (or a subset). Let $x_i = (x_{i1}, ..., x_{id})$ denote the realized value of the feature vector X_i and y_i be the corresponding class label.

There are several components in our combining method, called adaptive classification by mixing (ACM), including data splitting, the estimation of the conditional probability functions, and proper weighting. For ease in explanation, we describe the ACM method in three parts.

Let

$$f^1(x) = P(Y = 1 \mid X = x)$$

$$\cdots$$

$$f^m(x) = P(Y = m \mid X = x)$$

be the conditional probabilities of Y taking each label given the feature vector $X = x$ (they are referred to as the conditional probability functions). Let

$$f(x) = \left(f^1(x), f^2(x), \ldots, f^m(x)\right).$$

3.1 The main steps in combining the NN classifiers by ACM

We start with the individual NN classifiers. Let $\delta_{k,p}$ denote the k-NN classifier based on $Z^p =: (x_{i1}, ..., x_{ip}, y_i)_{i=1}^n$, where $1 \le p \le d$ is the number of variables used for the nearest neighbor procedure. In particular, $\delta_{1,1}$ is the 1-NN procedure using only the first variable, $\delta_{2,2}$ is the 2-NN procedure using the first two variables and so on.

Let $\hat{f}_{k,p}(x) = \hat{f}_{k,p}(x; Z^p)$ denote a certain estimator of f based on Z^p (the estimator will be given later in Section 3.2).

Let $\mathcal{J}$ and Ω be two subsets of the natural numbers that indicate which choices of the number of neighbors and the number of feature variables are considered. For each $k \in \mathcal{J}$ and $p \in \Omega$, we have a corresponding NN rule with k neighbors and the first p feature variables. For simplicity, we call the corresponding classifier k-NN-p classifier.

We propose the following algorithm for combining the k-NN-p classifiers with $k \in \mathcal{J}$ and $p \in \Omega$.

Algorithm ACM

Step 1. Obtain estimate $\hat{f}_{k,p}(x; Z^p) = \left(\hat{f}^1_{k,p}(x; Z^p), \hat{f}^2_{k,p}(x; Z^p), \ldots, \hat{f}^m_{k,p}(x; Z^p)\right)$ of $f(x)$ based on Z^p.

Step 2. Compute the weight $\widehat{w}_{k,p}$ for procedure $\delta_{k,p}$ (details to be given later in Section 3.3).

Step 3. The combined estimate of $f(x)$ is: for each $1 \le c \le m$,

$$\tilde{f}^c(x) = \sum_{p \in \Omega} \sum_{k \in \mathcal{J}} \widehat{w}_{k,p} \hat{f}^c_{k,p}(x; Z^p).$$

Step 4. Allocate x_{new} to group c if

$$\tilde{f}^c(x_{new}) = \text{the largest of } \left(\tilde{f}^1(x_{new}), \tilde{f}^2(x_{new}), ..., \tilde{f^m}(x_{new})\right).$$

When there are ties, one reasonable way to break them is to choose the label with the highest frequency in the data. If there are still ties, one can randomly pick a label among them.

Note that in our approach of combining, estimation of the conditional probability function f is needed. The combined estimate of f can be much more accurate than that based on selecting a single individual candidate. The essence of the weighting method is motivated from an information-theoretic consideration explained in Yang (2000).

3.2 Estimating the conditional probabilities

Here we provide the details of estimating the conditional probabilities in the ACM algorithm given in the previous subsection. Basically, the idea is to use the frequencies of the class labels in the neighborhood to estimate the conditional probabilities. Let M_1 be an integer.

Step 0. Randomly sample without replacement from $(x_i, y_i)_{i=1}^n$ to get a subset of size $n_1 = 2n/3$ (for simplicity, assume $2n/3$ is an integer). Let $Z^{(1)}$ denote this sub-sample.

Step 1. For each p, let $Z^{(p,1)}$ be the part of $Z^{(1)}$ using only the first p variables.

Step 2. For each x and $1 \le c \le m$, let

$$h_{k,p}^c(x; Z^{(p,1)}) = \text{ the number of class } c \text{ among the } k \text{ nearest neighbors in } Z^{(p,1)}.$$

Step 3. Repeat 0–2 $(M_1 - 1)$ more times and average $h_{k,p}^c(x; Z^{(p,1)})$ over the M_1 sub-samplings and let $\overline{h}_{k,p}^c(x)$ denote the average. Then let

$$\hat{h}_{k,p}^c(x) = \frac{\overline{h}_{k,p}^c(x) + 1}{k + m}$$

be the estimate of $f^c(x)$. Note that $\sum_{c=1}^m \hat{h}_{k,p}^c(x) = 1$ for each fixed pair (k, p).

In Step 3, the addition of 1 in the numerator of the formula that defines $\hat{h}_{k,p}^c(x)$ prevents the estimate of the conditional probability to be zero, and consequently avoids a problem in the weighting step for such a situation. Note that the denominator in the formula is correspondingly adjusted to ensure $\sum_{c=1}^m \hat{h}_{k,p}^c(x) = 1$.

3.3 Computing the weights

Let M_2 be an integer.

Step 1. Randomly permute the order of the observations. For each p, let $Z_E^p = (x_{i1}, x_{i2}, ..., x_{ip}, y_i)_{i=1}^{2n/3}$.

Step 2. Obtain the estimates of the conditional probabilities $\hat{h}_{k,p}^c(x)$ based on Z_E^p for each $k \in \mathcal{J}$ and $p \in \Omega$ (as described in the previous subsection).

Step 3. For each k, p, calculate

$$d_{k,p} = \prod_{i=2n/3+1}^{n} \left(\prod_{c=1}^{m} \hat{h}_{k,p}^c(x_i)^{I_{\{y_i = c\}}} \right),$$

where $I_{\{\}}$ denotes the indicator function.

Step 4. Compute the weight for the procedure $\delta_{k,p}$:

$$w_{k,p} = \frac{d_{k,p}}{\sum_{u\in\Omega}\sum_{l\in\mathcal{J}} d_{l,u}}.$$

Step 5. Repeat steps 1-4 $(M_2 - 1)$ more times and average the $w_{k,p}$ over the M_2 random permutations to obtain the final weight $\widehat{w}_{k,p}$.

Since $d_{k,p}$ depends on the order of the observations, the Step 5 eliminates this dependence when M_2 is large enough. The choice of the split proportion (2/3 for estimation and 1/3 for assessment) is based on our experience (from a rate of convergence stand point, any split proportion not going to 0 or ∞ would give the same rate of convergence). Note that pretending the estimates $\hat{h}^c_{k,p}(x)$ from the first part of the data (i.e., Z^p_E) are the correct conditional probability functions, $d_{k,p}$ is then the likelihood of the second part of the data (i.e., $Z^p_V = (x_{i1}, x_{i2}, ..., x_{ip}, y_i)^n_{i=2n/3+1}$) given the feature variables. The likelihood is simply a product of multinomial probabilities. Thus the weighting has an interpretation: if we put the uniform prior distribution on the candidate NN rules and pretend that the estimated conditional probability functions are the trues ones, then the weight $w_{k,p}$ is the posterior probability of $\delta_{k,p}$. In a formal sense, however, the ACM weighting is not a Bayes procedure. An advantage of our approach is that we do not need to deal with prior distribution assignment for parameters and it has a good theoretical risk property. For a theoretical risk bound of this weighting approach focusing on the 2-class case, see Yang (2000).

4 Data for our empirical study

In this section, we briefly describe the data sets used in our empirical study. We use the same three gene expression data sets studied in Dudoit *et al.* (2002). In general, a proper pre-processing of micro-array data is very important to facilitate a suitable data analysis. Since our focus in this work is on the comparison between CV selection and ACM combining, we simply follow the approach of Dudoit *et al.* (2002) for data pre-processing. See their paper for details.

4.1 Leukemia data set

This data was described in Golub *et al.* (1999). Gene expression levels were measured using Affymetrix high density oligonucleotide arrays with $p = 7,129$ human genes. There are 47 cases of acute lymphoblastic leukemia (ALL) (38 B-cell ALL and 9 T-cell ALL) and 25 cases of acute myeloid leukemia (AML). After the pre-processing, the date became the class labels and the 72×3571 matrix $X = x_{ij}$, where x_{ij} is the logarithm (base 10) of the expression level for gene j in mRNA samples i.

4.2 Lymphoma data set

Gene expression levels were measured using a specialized cDNA micro-array for studying the three most prevalent adult lymphoid malignancies: B-cell chronic lymphocytic leukemia (B-CLL), follicular lymphoma (FL), and diffuse large B-cell lymphoma (DLBCL) (Alizadeh *et al.* (2000)). This data set has gene expressions for $p = 4026$ genes in $n = 80$ mRNA samples. The mRNA samples were classified in three classes: 29 cases of B-CLL, 9 cases of FL, and 42 cases of DLBCL. The (i, j) entry of the 80×4026 matrix $X = x_{ij}$ is the logarithm (base 2) of CY5/CY3 fluorescence ratio for gene j in mRNA sample i.

4.3 NCI 60 data set

cDNA micro-arrays were used to measure the gene expression among 60 cell lines from the National Cancer Institute, which were derived from tumors with different sites of origin: 9 breast, 5 central nervous system (CNS), 7 colon, 8 leukemia, 8 melanoma, 9 non-small-cell-lung-carcinoma (NSCLS), 6

ovarian, 9 renal (Ross *et al.* (2000)). With data pre-processing, the 61×6830 matrix $X = x_{ij}$ has the entry of the base 2 logarithm of the Cy5/Cy3 fluorescence ratio for gene j in mRNA sample i.

We need to mention that we tried but could not access the data from the web sites provided in the Dudoit *et al.* (2002), but were able to obtain the data sets from other places. However, the leukemia and lymphoma data sets are slightly different: the number of observations in lymphoma data set is $n = 80$ ($n = 81$ in Dudoit *et al.* (2002)); and the leukemia data set has a different number of variables, but the number of variables became the same after the data pre-processing.

5 Study Design

As is commonly used in an empirical study in pattern recognition, we compare the different approaches of classification by data splitting: each data set is randomly split into a learning set (LS) and a test set (TS). The LS is used to build the classifiers, i.e., the classifier based on CV selections and the combined classifier based on ACM in our context. The TS is used to compare the accuracy of the classifiers. Note that the LS needs to be further split for both CV selection and ACM weighting. To eliminate the influence of the order of the observations, we replicate the process 150 times by randomly permuting the order of the observations.

Since our data have relatively small sample sizes, and our main purpose is to compare combining classifiers with selecting a single winner, as opposed to estimating the generalization error rates, we choose the test size to be one third of the data (2:1 scheme) instead of a smaller proportion for the TS commonly seen in the machine learning literature.

In our experiments, we found that the choice of p, the number of variables (genes), is an important issue. When the candidate values of p are widely separated, CV basically has little difficulty finding the best combination of the neighbor size k and the number of genes p. In a real application, however, a sparse choice of p looks ad hoc and may be too rough for achieving a good classification accuracy. There are two reasonable approaches to address this issue. One is to consider all the numbers of genes up to an upper bound (say 300). For saving computation time, one can consider equally spaced integers in the range with a reasonably small spacing. Another approach is to do some preliminary analysis to have a sense about which p's are good candidates for best performance of classification and then restrict attention to them.

We describe some specifics of the methods in competition below.

Combining method We choose $M_1 = 100$ and $M_2 = 10$ (recall that M_1 and M_2 are the numbers of random sampling/permutation for estimating the conditional probability and computing the weights respectively).

Cross-validation We consider three CV schemes. In cases of 2-fold and 3-fold cross-validation (1:1 and 2:1 scheme), the best k neighbor and variable sizes are selected by choosing the lowest average error rates over $M = 50$ replications of permuting the observations before splitting the learning set). The other CV scheme is the familiar *leave-one-out* CV.

Let us comment on the difference between our study design and the corresponding part in Dudoit *et al.* (2002). In their paper, the number of neighbors considered is in the range $1 \leq k \leq 21$ and is selected to minimize the test error rate by the *leave-one-out* cross-validation. Their interest was in selecting the number of neighbors alone and the number of variables was not involved in the CV selection (it was fixed). In contrast, in this work both p and k are involved in the selection and combining methods. Note that p can have a significant influence on the classification accuracy for two of the data sets and the effect is even larger than that of k in the relevant range.

Finally, let us briefly discuss about the computation aspects. Our programs were written using **R** and **C** (the complied C code was linked to **R** for speeding up the computation). With our current

programs, combining procedures takes substantially more time (even twice sometimes) compared to the cross-validation parts.

6 Results

6.1 The effects of p and k on classification accuracy

Obviously, with the number of genes being so large for these data sets, considering all possible combinations of the neighbor size and the number of genes (variables) is unwise. It is a good idea to first have a reasonable understanding on which p and k values are potentially useful to be considered for combining or selection.

Here we briefly summarize the effects of p and k on classification accuracy of the corresponding NN rules for the three data sets.

For all the three data sets, using more than 10 neighbors does not seem to be advantageous and can even hurt the performance. For the leukemia data set, the increase of p from 10 to 200 or bigger does not have a dramatic effect. However, in cases of the lymphoma and NCI60 data sets, the choices of p around 200 give substantially better accuracy than small values of p. Note also that the accuracy of the nearest neighbor classifiers is much worse for NCI60 data than for the other two cases.

6.2 Combining improves over cross-validation

The purpose of this subsection is to demonstrate the effectiveness of the combining method compared with cross-validations.

Given in Figures 1-3 are the box-plots of the test error rates of the competing classifiers for the data sets based on 150 random splits of each data set into LS and TS. The figures use acronyms in listing the classifiers:

com The combining procedure ACM as described in Section 3.

cross1 Cross-Validation (3-fold, 2:1 scheme).

cross2 Cross-Validation (2-fold, 1:1 scheme).

cross3 Cross-Validation (*leave-one-out*).

In all these experiments, the number of neighbors is given the choices from 1 to 10 (i.e, $\mathcal{J} = \{1, 2, ..., 10\}$). For the leukemia data, p has the choices of 7, 9, 11, 13, 15; for the lymphoma data, p has the choices of 180, 190, 200, 210, 220; and for the NCI60 data, p has the choices of 140, 160, 180, 200, 220. Note that these choice of p are respectively in the ranges that seem to give good performance in the preliminary analysis.

From the box-plots, clearly combining by ACM significantly improves the performance over all the CV methods. For the leukemia data, the mean error rate was reduced by at least 39% by the combining procedure from 5.4%, 5.4% and 5.8% of the three CV methods to 3.3%. Note that the median error rates are the same (4.2%) for all the competing methods. For the lymphoma data, with ACM combining, the mean error rate was reduced by at least 26% from 4.3%, 4.1% and 4.4% to 3.0%. Interestingly, the median error rate was reduced from 3.7%, 3.7%, and 3.7% to 0%. For the NCI60 data, ACM reduced the mean error rate by at least 19% from 24.6%, 24.7% and 23.2% to 18.7%. The median error rate was reduced similarly. All of the reductions mentioned above were statistically significant.

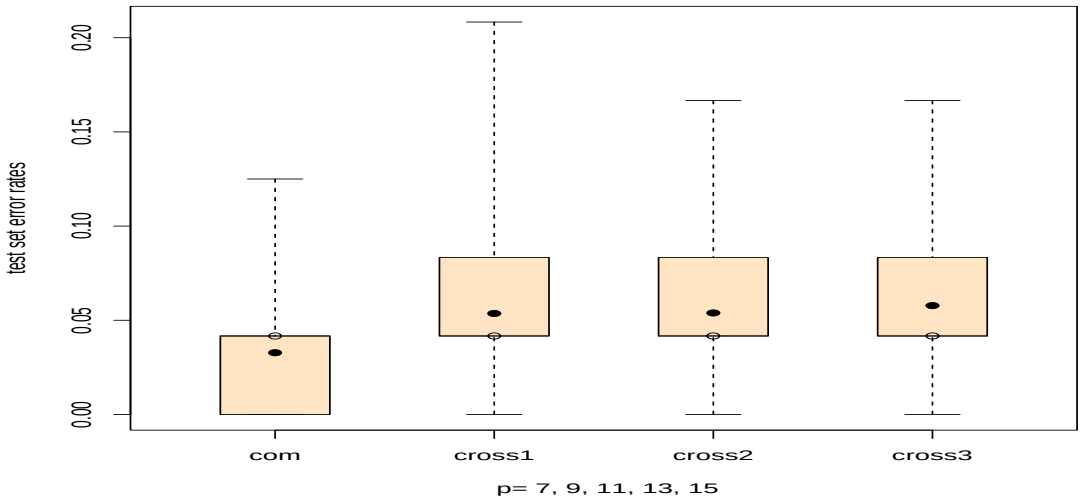

Figure 1: *Leukemia data.* Box-plots of test error rates for the combining and selection procedures based on the NN rules with $p = 7, 9, 11, 13, 15$ and $k = 1, ..., 10$; $N = 150$ LS/TS random splits with 2:1 sampling scheme

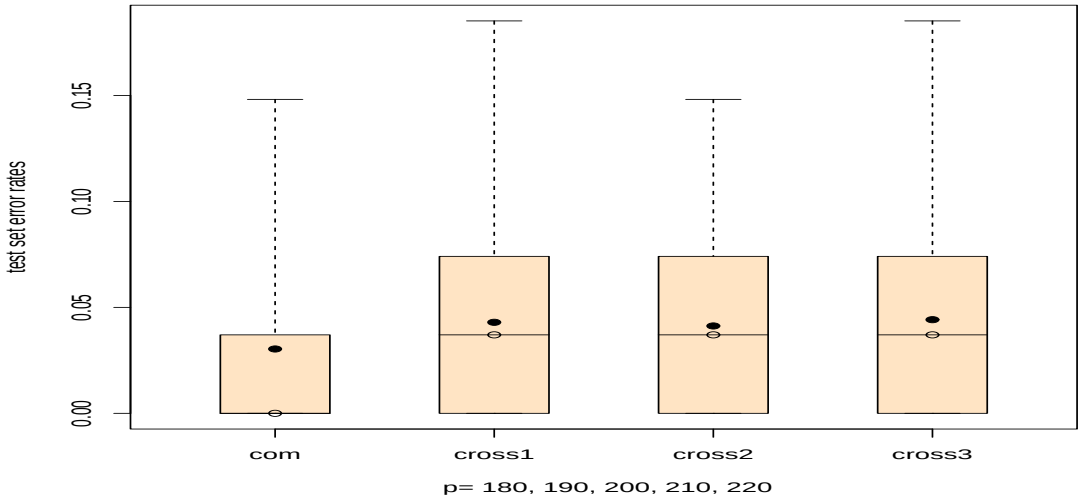

Figure 2: *Lymphoma data.* Box-plots of test error rates for the combining and selection procedures based on the NN rules with $p = 180, 190, 200, 210, 220$ and $k = 1, ..., 10$; $N = 150$ LS/TS random splits with 2:1 sampling scheme

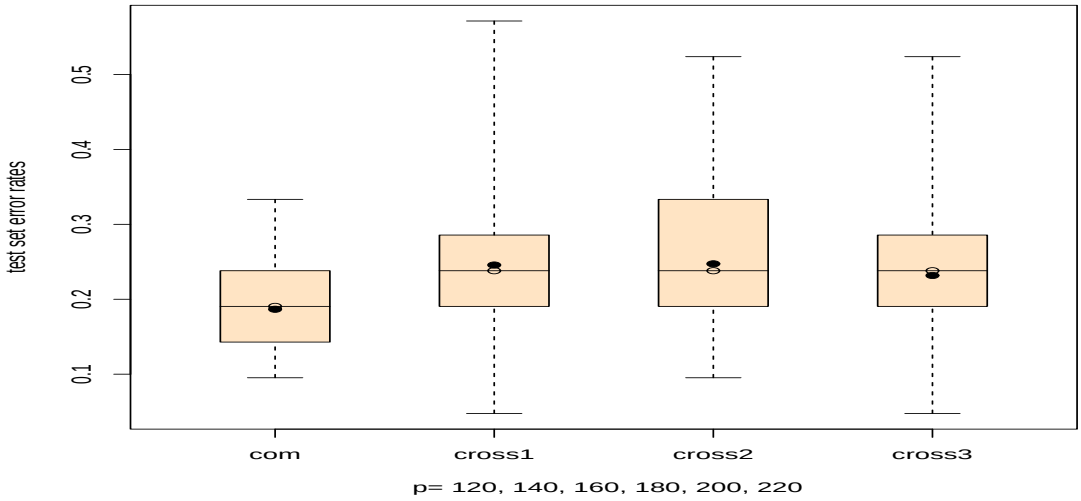

Figure 3: *NCI60 data*. Box-plots of test error rates for the combining and selection procedures based on the NN rules with $p = 120, 140, 160, 180, 200, 220$ and $k = 1, ..., 10$; $N = 150$ LS/TS random splits with 2:1 sampling scheme

6.3 When does combining work better than selection?

In the previous subsection, we demonstrated the advantage of combining NN rules by ACM over CV selections in terms of classification error rate. However, it is better not to stop at the fact that combining *can* work better than selection. It is desirable to gain more insight for a better understanding on the comparison between selection and combining. Is combining NN rules generally better than selecting a single candidate? If not, when does combining offer improvement?

These questions are important for applications and proper answers can help us to achieve better classification accuracy.

We give another three examples, one for each data set. Here we fix k to be 1 (i.e., consider only 1-NN rules) and consider $p = 10$, 40, 100, 150, 200 for the leukemia data, $p = 100$, 200, 300 for the lymphoma data and $p = 120$, 160, 200 for the NCI60 data sets. The box-plots of the test error rates are presented in Figures 4-6. From these figures, combining does not always improve and can perform worse than the CV selections.

Therefore, it becomes clear that it is not necessarily a good idea to always combine candidate classifiers. But how do we know when to combine the candidate classifiers and when to select one of them?

A key issue to address the above question is whether the CV selection methods are having difficulty or not. Roughly speaking, if the CV methods are highly unstable, combining has a great potential to improve. It should be noted that for Figures 4-6, the candidate values of p are much more sparse compared to those for Figures 1-3, respectively. Thus it is possible that the CV selections are much easier for the latter three figures. More formally, in our setting, we can examine the frequencies that the allowed combinations of k and p are selected by the CV methods as the winner over the 150 random splits of the original data. If the frequencies are highly concentrated, it indicates that the CV methods are quite stable. Otherwise, the selection is unstable.

Tables 1-6 give the frequencies for the 6 examples. They indeed support the aforementioned view on the relative performance between combining and selection. For the first three examples where ACM combining is seen to have a substantial advantage, the frequencies are well spread out. For the latter three examples, for both the leukemia data and the NCI60 data, the majority of the counts are in one cell (or

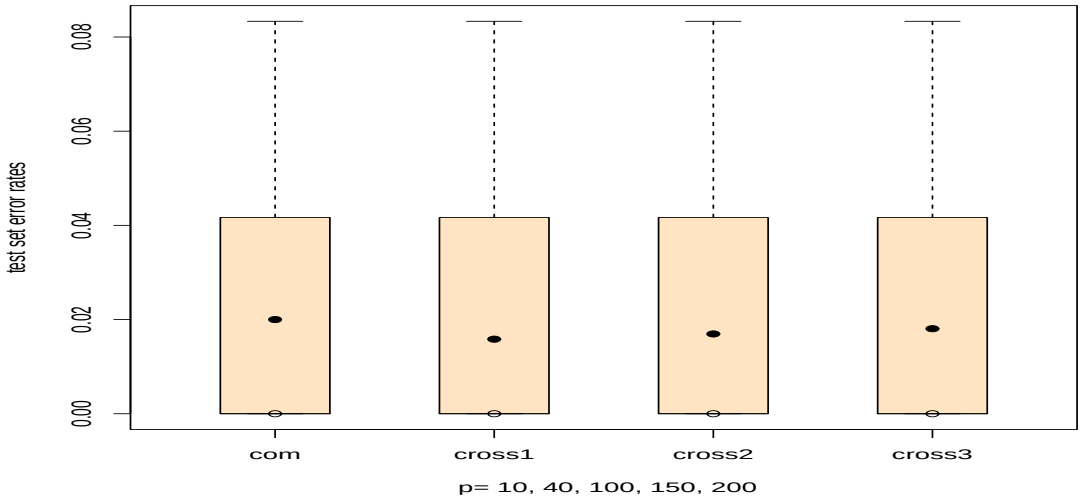

Figure 4: *Leukemia data.* Box-plots of test error rates for the combining and selection procedures based on the NN rules with $p = 10, 40, 100, 150, 200$ and $k = 1$; $N = 150$ LS/TS random splits with 2:1 sampling scheme

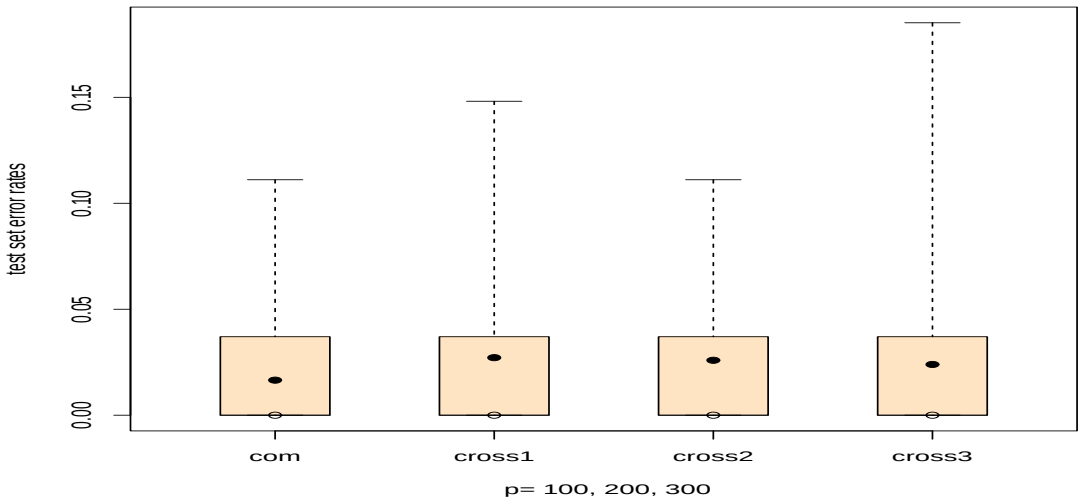

Figure 5: *Lymphoma data.* Box-plots of test error rates for the combining and selection procedures based on the NN rules with $p = 100, 200, 300$ and $k = 1$; $N = 150$ LS/TS random splits with 2:1 sampling scheme

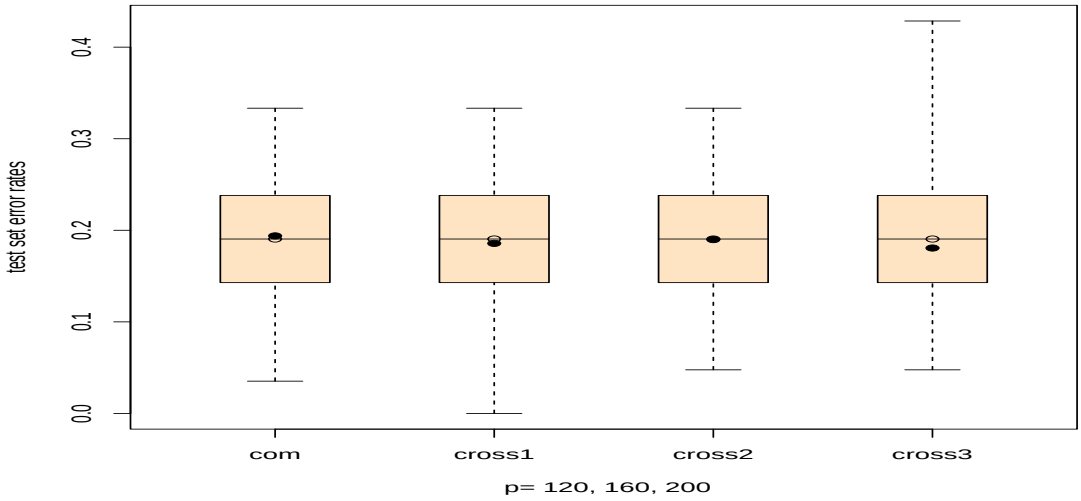

Figure 6: *NCI60 data.* Box-plots of test error rates for the combining and selection procedures based on the NN rules with $p = 120, 160, 200$ and $k = 1$; $N = 150$ LS/TS random splits with 2:1 sampling scheme

almost so in one case) and thus the selection process is relatively stable and correspondingly, combining performs no better; for the lymphoma data, the frequencies are less concentrated and combining is slightly advantageous.

In summary, the above study supports that when the candidate classifiers are hard to be distinguished, compared to averaging them properly, selecting a single "winner" can bring in much larger variability in the classifier. On the other hand, when one classifier is clearly better, averaging with poor ones can damage the performance (unless their assigned weight are small enough). Therefore combining classifiers is advantageous to selecting a single candidate when there is a certain degree of uncertainty in choosing the best one. In general, combining and selecting are both useful, depending on the situation.

From above, in real applications, we recommend that one assesses the selection instability (e.g., via the selection frequency table). If there is clearly little instability, there is not much incentive to try combining. On the other hand, when the selection methods are highly unstable, one should not be overly confident about the selected classifier and combining should be seriously considered.

7 Summary and Discussion

Nearest neighbor methods are widely used in pattern recognition. In the context of tumor classification with micro-array gene expression data, Dudoit *et al.* (2002) concluded that "simple classifiers such as DLDA and NN performed remarkably well compared to more sophisticated ones such as aggregated classification trees". In this work, we focused on the NN classifiers and addressed the practically important issue of the choice between selecting and combining NN classifiers of different combinations of the neighbor size and the number of feature variables. We proposed a combining method ACM and empirically studied its performance relative to three cross-validation methods (3-fold, 2-fold and delete-one). In addition, an effort was made to gain insight on when combining candidate classifiers is advantageous to selecting one of them.

The results showed that ACM can substantially improve the classification accuracy over the CV methods even up to 39%. However, this is not always the case. In fact, the additional examples showed that when the CV methods can pretty much easily identify the best classifier among the candidates, combining does not help and can even hurt the performance. Thus model/procedure selection and

		k=1	2	3	4	5	6	7	8	9	10
cross1	p=7	45	5	5	3	7	1	1	4	1	3
	p=9	0	1	4	1	3	3	5	0	4	2
	p=11	2	1	2	1	3	2	0	3	2	1
	p=13	3	0	4	1	5	1	6	1	1	1
	p=15	3	2	1	4	1	0	1	0	0	0
cross2	p=7	26	10	13	4	3	4	3	3	6	1
	p=9	4	5	5	1	4	1	1	4	1	2
	p=11	1	2	0	1	2	1	4	4	2	0
	p=13	5	3	8	2	6	1	1	1	0	0
	p=15	1	0	1	2	1	0	0	0	0	0
cross3	p=7	47	2	4	4	3	3	2	3	0	3
	p=9	1	1	2	1	2	5	4	1	3	2
	p=11	2	2	2	0	2	1	1	3	3	5
	p=13	6	5	0	3	3	0	1	1	2	1
	p=15	2	2	3	3	2	0	2	0	0	0

Table 1: *Leukemia data.* Frequencies of the NN classifiers being selected with $p = 7, 9, 11, 13, 15$ and $k = 1, 2, ..., 10$; $N = 150$ random splits with 2:1 sampling scheme

		k=1	2	3	4	5	6	7	8	9	10
cross1	p=180	19	7	9	7	6	3	8	8	4	6
	p=190	21	4	6	0	1	1	2	0	1	0
	p=200	11	0	3	0	0	0	0	0	0	1
	p=210	11	0	0	1	1	0	0	0	0	1
	p=220	6	1	1	0	0	0	0	0	0	0
cross2	p=180	12	10	10	14	4	9	8	11	12	10
	p=190	18	1	2	1	1	1	2	2	0	0
	p=200	8	0	0	0	1	1	0	0	0	1
	p=210	8	0	0	0	0	0	0	0	0	0
	p=220	2	0	0	0	1	0	0	0	0	0
cross3	p=180	13	3	4	1	2	3	1	1	2	1
	p=190	8	2	6	4	10	8	3	4	0	0
	p=200	11	8	0	1	2	0	3	1	0	0
	p=210	20	6	0	2	0	0	0	2	0	0
	p=220	14	2	0	1	1	0	0	0	0	0

Table 2: *Lymphoma data.* Frequencies of the NN classifiers being selected with $p = 180, 190, 200, 210, 220$ and $k = 1, 2, ..., 10$; $N = 150$ random splits with 2:1 sampling scheme

		k=1	2	3	4	5	6	7	8	9	10
cross1	p=120	28	8	6	4	9	9	9	5	2	7
	p=140	7	2	2	1	2	1	3	1	3	3
	p=160	4	0	2	3	1	0	1	2	3	1
	p=180	5	0	1	0	0	0	0	1	2	1
	p=200	1	1	1	1	0	0	0	0	0	0
	p=220	4	0	1	0	0	0	0	0	0	0
cross2	p=120	26	3	7	10	7	4	3	3	7	7
	p=140	2	2	2	4	6	1	2	2	5	3
	p=160	5	5	4	2	3	2	3	3	1	3
	p=180	3	2	0	2	0	0	0	0	0	0
	p=200	0	0	0	0	0	0	0	1	2	1
	p=220	2	0	0	0	0	0	0	0	0	0
cross3	p=120	23	4	5	4	3	1	0	6	4	1
	p=140	11	3	2	3	3	0	2	4	1	2
	p=160	14	5	3	1	0	3	0	0	1	0
	p=180	15	3	2	1	0	0	0	0	1	0
	p=200	2	3	3	1	0	0	0	0	0	1
	p=220	1	4	2	2	0	0	0	0	0	0

Table 3: *NCI60 data.* Frequencies of the NN classifiers being selected with $p = 120, 140, 160, 180, 200, 220$ and $k = 1, 2, ..., 10$; $N = 150$ random splits with 2:1 sampling scheme

	cross1	cross2	cross3
p=10	96	109	74
p=40	26	27	47
p=100	11	14	29
p=150	11	0	0
p=200	6	0	0

Table 4: *Leukemia data.* Frequencies of the NN classifiers being selected with $p = 10, 40, 100, 150, 200$ and $k = 1$; $N = 150$ random splits with 2:1 sampling scheme

	cross1	cross2	cross3
p=100	78	75	48
p=200	45	37	57
p=300	27	38	45

Table 5: *Lymphoma data.* Frequencies of the NN classifiers being selected with $p = 100, 200, 300$ and $k = 1$; $N = 150$ random splits with 2:1 sampling scheme

	cross1	cross2	cross3
p=120	95	83	79
p=160	46	52	50
p=200	9	15	21

Table 6: *NCI60 data.* Frequencies of the NN classifiers being selected with $p = 120, 160, 200$ and $k = 1, 2, ..., 10$; $N = 150$ random splits with 2:1 sampling scheme

combining both have their places in data analysis. In application, one can do some preliminary analysis (e.g., via data splitting and testing) to have a reasonable sense about the effects of the neighbor size and the number of feature variables. Instability of a CV method can be reasonably reflected by the distribution of the frequencies of the selected neighbor size and the number of feature variables over a number of random splits of the data and testing. Such analysis and the information gained are very helpful to make a wise decision on selecting or combining the classifiers for a better accuracy.

Our study did not include other popular classifiers from statistics and machine learning such as linear discriminant analysis (Fisher (1936)), CART (Breiman *et al.* (1984)), neural networks (e.g., Ripley (1996)), support vector machines (Vapnik (2000)) and so on. There are two reasons for this. One is that the results of Dudoit *et al.* (2000) suggest that for these gene expression data sets, the other procedures do not have advantage; and the other is that focusing on the nearest neighbor methods gives a more clear picture for comparing the combining and selection methods. Nonetheless, for other types of data, those classifiers can be advantageous and it is then worthwhile to include them as candidate classifiers.

8 Acknowledgments

The authors thank Professor Dan Nettleton for helpful comments. The paper benefited from the second author's visit to the Institute for Mathematics and its Applications (IMA) at University of Minnesota as a New Direction Visiting Professor in 2003-2004. The work of the second author was also partly supported by NSF CAREER grant DMS0094323.

References

[1] Alizadeh, A.A., Eisen, M.B., Davis, R.E., MA, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., YU, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Jr, J.H., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O., and Staudt, L.M. (2000) Different types of diffuse large b-cell lymphoma identified by gene expression profiling, *Nature*, **403**, 503-511.

[2] Allen, D.M. (1974) The relationship between variable selection and data augmentation and a method of prediction, *Technometrics*, **16**, 125-127.

[3] Ambroise, C. and McLachlan, G. (2002) Selection bias in gene extraction in tumour classification on basis of microarray gene expression data, *PNAS*, **99**(10), 6562-6566.

[4] Breiman, L., Friedman, J.H., Olshen, R. and Stone, C.J. (1984) *Classification and Regression Trees*, Wadsworth, Belmont, CA.

[5] Breiman, L. (1996) Bagging predictors, *Machine Learning*, **24**, 123-140.

[6] Cover, T.M. and Hart, P.E. (1967) Nearest neighbor pattern classification, *IEEE*, (Transactions on Information Theory, **13**, 21-27.).).

[7] Devroye, L., Györfi, L., and Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag Inc, New York.

[8] Dudoit, S., Fridlyand, J. and Speed, T. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the America Statistical Association*, **97**(457), 77-87.

[9] Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, **7**, 179-188,

[10] Fix, E. and Hodges, J.L. (1951) Discriminatory analysis, nonparametric discrimination, consistency properties, Technical Report 4. Randolph Field, TX: United States Air Force, School of Aviation Medicine, 261-279.

[11] Freund, Y. and Schapire, R.E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, **55**(1), 119-139.

[12] Geisser, S. (1975). The predictive sample reuse method with applications, *Journal of the American Statistical Association*, **70**(350), 320-328.

[13] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, **286**, 531-537.

[14] Hastie, T., Tibshirani, R., and Buja, A. (1994) Flexible discriminant analysis, *Journal of the American Statistical Association*, **89**, 1255-1270.

[15] Ripley, B.D. (1996) *Pattern Recognition and Neural Networks*, Cambridge, U.K: Cambridge University Press.

[16] Ross, D.T., Scherf, U., Eisen, M.b., Perou, C.M., Spellman, P., Iyer, V., Jeffrey, S.S., de Rijn, M.V., Pergamenschikov, A., Lee, J.C.F., Lashkari, D., Myers, T.G., Weinstein, J.N., Bostein, D., and Brown, P.O. (2000) Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genetics*, **24**, 227-234.

[17] Speed, T.G. (2003) *Statistical Analysis of Gene Expression Microarray Data*, CRC Press.

[18] Stone, M. (1974) Cross-validation choices and assessment of statistical predictions, *J. Roy. Statist. Soc Ser* B, **36**, 111–147.

[19] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression, *PNAS*, **99**, 6567-6572.

[20] Vapnik, V. (1982) *Estimation of Dependencies Based on Empirical Data*, Springer-Verlag Inc, New York. (English translation from Russian: *Nauka*, Moscow, 1979.)

[21] Vapnik, V. (1998) *Statistical Learning Theory*, John Wiley, New York.

[22] West, M. *et al.* (2001) Predicting the clinical status of human breast cancer using gene expression profiles, *Proc. Natl. Acad. Sci.*, 98:11462-11467.

[23] Yang, Y. (2000) Adaptive estimation in pattern recognition by combining different procedures, *Statistica Sinica*, **10**, 1069-1089.

[24] Yang, Y. (2003) Regression with multiple candidate models: selecting or mixing?, *Statistica Sinica*, **13**, 783-809.